Understanding the Generalization Benefit of Model Invariance from a Data Perspective

Anonymous Author(s)
Affiliation
Address
email

Abstract

Machine learning models that are developed to be invariant under certain types of data transformations have shown improved generalization in practice. However, a principled understanding of why invariance benefits generalization is limited. Given a task, there is often no principled way to select "suitable" data transformations under which model invariance guarantees better generalization. This paper understands the generalization benefit of model invariance by introducing the *sample cover induced by transformations*, i.e., a representative subset of a dataset that can approximately recover the whole dataset using transformations. For any data transformations, we provide refined generalization bounds for invariant models based on the sample cover. We also characterize the "suitability" of a set of data transformations by the *sample covering number induced by transformations*, i.e., the smallest size of its induced sample covers. We show that we could tighten the generalization bounds for "suitable" transformations that have a small sample covering number. In addition, our proposed sample covering number can be empirically evaluated and thus provides principled guidance for selecting transformations to develop model invariance for better generalization. In experiments on CIFAR-10 and a 3D dataset, we empirically evaluate sample covering numbers for some commonly used transformations and show that the smaller sample covering number for a set of transformations (e.g. the 3D-view transformation) indicates a smaller gap between the test and training error for invariant models, which verifies our propositions.

1 Introduction

Invariance is ubiquitous in many real-world problems. For instance, categorical classification of visual objects is invariant to slight viewpoint changes [14, 2, 17], text understanding is invariant to synonym substitution and minor typos [36, 25, 19]. Intuitively, models capturing the underlying invariance exhibit improved generalization in practice [16, 11, 33, 12, 30, 12]. Such generalization benefit is especially crucial when the data are scarce as in some medical tasks [31], or when the task requires more data than usual as in cases of distribution shift [26] and adversarial attack [28, 32, 5].

A commonly accepted intuition attributes the generalization benefit of model invariance to the reduced model complexity, especially the reduced sensitivity to spurious features. However, a principled understanding of why model invariance helps generalization remains elusive, thus leaving many open questions. Since model invariance may come at a cost (e.g., compromised accuracy, increase computational overhead), given a task, how should we choose among various data transformations under which model invariance guarantees better generalization? If existing data transformations are not good enough for a given task, what is the guiding principle to find new ones? The lack of a principled understanding limits better exploitation of model invariance to further improve

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

generalization. In addition, since identifying instructive generalization bound is a central topic in machine learning, we may expect to refine the existing generalization bounds by additionally considering the model invariance.

The many faces of data transformations and model classes pose significant challenges to a principled understanding of model invariance's generalization benefit. To address this, [2, 4, 23, 29, 27] characterize the input space and show that certain data transformations equivalently shrink the input space for invariant models, which then simplify the input and improves generalization. From another perspective, [15, 21] directly characterize the function space and show that the volume of the invariant model class is reduced, which then simplifies the learning problem and improves generalization. These understandings provide valuable insights, yet they may suffer from the high-dimensionality of input data or require the model invariance to be obtained exclusively via feature averaging. Some certain assumptions on data transformations (e.g., finiteness, group structure with certain measure) also make these understandings less applicable to more general data transformations.

In this paper, we understand the generalization benefit of model invariance by characterizing sample-dependent properties of data transformations and then provide practical guidance for selecting suitable data transformations. Our work motivates exploring more suitable data transformations to better improve invariant models' generalization performance.

1.1 Summary of Contributions

Data-perspective understanding. (Section 3) We understand the generalization benefit of model invariance via characterizing the sample-dependent properties of data transformations. We introduce the notion of sample cover induced by data transformations and establish connections between the small sample covering number of data transformations and the small generalization bound of invariant models. Since this general connection requires a strong assumption on data transformations to be instructive, we relax the assumption by further assuming the Lipschitzness of model class to get a refined generalization bound. To understand the generalization benefit of model invariance in a more interpretable way, we also consider linear model class and show that the model complexity of invariant models can be much smaller than its counterpart, depending on the sample-dependent properties of data transformations.

Guidance for data transformation selection. (Section 4) To guide the data transformation selection in practice, we propose to use the sample covering number induced by data transformations as a suitability measurement to estimate the generalization benefit for invariant models. This measurement is model-agnostic and applies generally to any data transformation. We also introduce an algorithm to estimate the sample covering number in practice.

Empirical evaluations. (Section 5) We empirically estimate the sample covering number for some common data transformations on CIFAR-10 and ShapeNet (a 3D-dataset) to evaluate their potential generalization benefit. The 3D-view transformation renders smaller empirical sample covering number than others by a large margin on ShapeNet, while cropping is the most favorable among others. Then, we do invariant model training under those data transformations by data-augmentation and invariance loss penalization to evaluate the actual generalization benefit for invariant models. Results show clear correlation between smaller sample covering numbers of data transformations and better generalization benefit for invariant models.

2 Preliminaries

Data transformations. We refer to the data transformation as a function mapping from the input space $\mathcal{X} \to \mathcal{X}$, and data transformation*s* as a set of such functions. We do not assume data transformations to have group structures as many non-invertible transformations (e.g., cropping) does not fit into a group structure directly. For a set of data transformations $\mathcal{G} = \{g : \mathcal{X} \to \mathcal{X}\}$ and a data point $\boldsymbol{x} \in \mathcal{X}$, we overload the notion of orbit in group theory and denote by $\mathcal{O}_{\mathcal{G}}(\boldsymbol{x})$ the orbit of $\boldsymbol{x}$ defined as follows. The *orbit* $\mathcal{O}_{\mathcal{G}}(\boldsymbol{x})$ of a data point $\boldsymbol{x}$ w.r.t. data transformations $\mathcal{G}$ is the collection of outputs after any transformation $g \in \mathcal{G}$ on $\boldsymbol{x}$: $\mathcal{O}_{\mathcal{G}}(x) = \{g(\boldsymbol{x}) \in \mathcal{X} : g \in \mathcal{G}\}$.

Model invariance. A model $h : \mathcal{X} \to \mathcal{Y}$ is said to be *invariant* under data transformations $\mathcal{G}$ if $h(g(\boldsymbol{x})) = h(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$ and any $g \in \mathcal{G}$. We refer to a class of such invariant models as the *$\mathcal{G}$-invariant model class*.

Complexity measurements. *Covering number* [22] and *Rademacher complexity* [22] are two commonly used complexity measurements for model classes (including neural networks [6]) that can provide uniform generalization bound. The covering number can also be directly used to upper bound the Rademacher complexity via Dudley's theorem [13, 22].

Covering number. Let $(\mathcal{F}, d)$ be a (pseudo)metric space with some (pseudo)metric[1] d. An *ϵ-cover* of a set $\mathcal{H} \subseteq \mathcal{F}$ is defined as a subset $\widehat{\mathcal{H}} \subseteq \mathcal{H}$ such that for any $h \in \mathcal{H}$, there exists $\widehat{h} \in \widehat{\mathcal{H}}$ such that $d(h, \widehat{h}) \leq \epsilon$. The covering number $N(\mathcal{H}, \epsilon, d)$ is defined as the minimum cardinality of an ϵ-cover (among all ϵ-covers) of $\mathcal{H}$. In this paper, we use the concept of covering number both on model class and on dataset.

Empirical Rademacher complexity. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathbb{R}$. Given a sample $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$, the *empirical Rademacher complexity* of model class $\mathcal{H}$ is defined as: $\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\right]$ where $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_n]^\top$ is the vector of i.i.d. Rademacher random variables, each uniformly chosen from $\{-1, 1\}$.

Generalization gap. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample drawn i.i.d. from data distribution $\mathcal{D}$, and $\mathcal{H}$ be a model class. Given a loss function $\ell : \mathbb{R} \to [0, 1]$, for a $h \in \mathcal{H}$, we define the *empirical error* as $R_{\mathcal{S}}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(\boldsymbol{x}_i), y_i)$, the *generalization error* as $R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(h(\boldsymbol{x}), y)]$, and the *generalization gap* as $R(h) - R_{\mathcal{S}}(h)$. In this paper, we refer to a small generalization gap as the good generalization. The covering numbers and empirical Rademacher complexity of a model class can both be used to give an uniform bound of the models' generalization gap [22] which we include in the Appendix A for reference.

3 Understanding the Generalization Benefit of Model Invariance

In this section, we understand the generalization benefit of model invariance by refining generalization bounds for invariant models. Instead of directly characterizing the function space of invariant models, we first characterize how data transformations affect samples in a model-agnostic way, and then establish a connection with the model's generalization gap. Albeit less elegant, this analysis is more applicable regardless of how the invariance is obtained and is more tractable for empirical evaluation.

We first introduce the notion of *sample cover induced by data transformations*, which is a representative subset of examples induced by the data transformations. The intuition is that invariant models may have consistent behavior on a data point and its approximation in the sample cover so that we only need to consider the invariant models' behavior on the sample cover.

The definition of sample cover relies on a pseudometric induced by the data transformations $\mathcal{G}$. Let $\|\cdot\|$ be any norm on the input space $\mathcal{X}$. Given the set of transformations $\mathcal{G}$, we define the pseudometric[2]

$$\rho_{\mathcal{G}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \inf_{\gamma \in \Gamma(\boldsymbol{x}_1, \boldsymbol{x}_2)} \int_\gamma c(r) dr, \quad \text{where } c(r) = \begin{cases} 0, & \text{if } r \text{ is in some orbit} \\ \|r\|, & \text{otherwise} \end{cases}$$

where Γ denotes the set of all paths from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$.

Definition 3.1 (Sample cover induced by data transformations)**.** Let $(\mathcal{X}, \rho_{\mathcal{G}})$ be a pseudometric space and $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. An *ϵ-sample cover* $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ of the sample $\mathcal{S}$ induced by data transformations $\mathcal{G}$ at scale ϵ is defined as a representative subset of the sample $\mathcal{S}$ such that for any $\boldsymbol{x} \in \mathcal{S}$, there exists $\widehat{\boldsymbol{x}} \in \widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ such that $\rho_{\mathcal{G}}(\boldsymbol{x}, \widehat{\boldsymbol{x}}) \leq \epsilon$.

Definition 3.2 (Sample covering number induced by data transformations)**.** The *sample covering number* $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by data transformations $\mathcal{G}$ is defined as the minimum cardinality of an ϵ-sample cover:

$$N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}}) = \min\{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}| : \widehat{\mathcal{S}}_{\mathcal{G},\epsilon} \text{ is an } \epsilon\text{-sample cover of } \mathcal{S}\}. \tag{3.1}$$

[1] A pseudometric is a metric if and only if it separates distinct points, namely $d(x, y) > 0$ for any $x \neq y$.

[2] Note that $\rho_{\mathcal{G}}$ is not a metric since it allows $\rho_{\mathcal{G}}(x, y) = 0$ for $x \neq y$.

The sample cover induced by some data transformations $\mathcal{G}$ enables us to characterize the $\mathcal{G}$-invariant models' complexity by only considering models' behaviour on the sample cover. Moreover, the sample covering number induced by $\mathcal{G}$ directly leads to the following model class complexity bound which holds for any $\mathcal{G}$-invariant model class.

Proposition 3.3. Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}$ be a model class that has bounded output and is invariant to $\mathcal{G}$. Then the following Rademacher upper-bound for $\mathcal{H}$ holds up to some constant.

$$\mathfrak{R}_S(\mathcal{H}) \lesssim O\left(\sqrt{\frac{N(0, \mathcal{S}, \rho_{\mathcal{G}})}{n}}\right). \tag{3.2}$$

Proposition 3.3 shows that any $\mathcal{G}$-invariant model class has a bounded empirical Rademacher complexity w.r.t. the sample covering number $N(0, \mathcal{S}, \rho_{\mathcal{G}})$ induced by $\mathcal{G}$. Smaller sample covering number thus gives tighter model class complexity bound and consequently tighter generalization bound (see Appendix A).

However, Proposition 3.3 is instructive only when the data transformations yield $N(0, \mathcal{S}, \rho_{\mathcal{G}}) \ll n$ on the sample $\mathcal{S}$, which means we can exactly recover all data points in $\mathcal{S}$ using a small-sized subset and data transformations $\mathcal{G}$. This condition is typically not satisfied by data transformations that can only generate an orbit with measure zero (w.r.t. data measure) at most data points. For example, the rotation transformations on CIFAR-10 do not satisfy this condition, since no two images in CIFAR-10 are rotated versions of each other. To relax this restriction on data transformations to better understand the generalization benefit brought by more commonly used data transformations (e.g., rotation), we further consider specific model class properties which may expand the data point orbit so as to get finer results. We study the Lipschitz model class in section 3.1 as well as the linear model class for a sharper result in section 3.2.

3.1 Special Case: Lipschitz Model Class

This subsection shows the generalization benefit brought by any data transformations by considering the Lipschitz model class. The following theorem provides a refined model complexity result for the invariant Lipschitz model class.

Theorem 3.4 (Refined model covering number of $\mathcal{G}$-invariant $\mathcal{H}^{\mathsf{Lip}}$)**.** Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}^{\mathsf{Lip}}$ be a model class such that every $h \in \mathcal{H}^{\mathsf{Lip}}$ is κ-Lipschitz with respect to $\|\cdot\|$ (which is used in defining the sample cover) and invariant to $\mathcal{G}$. Then

$$N\big(\tau, \mathcal{H}^{\mathsf{Lip}}, L_2(\mathbb{P}_{\mathcal{S}})\big) \leq \inf_{\epsilon \geq 0, \widehat{\mathcal{S}}_{\mathcal{G},\epsilon}} N\Big(\tau - \kappa\epsilon\sqrt{1 - \frac{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}|}{n}}, \mathcal{H}^{\mathsf{Lip}}, L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})\Big), \tag{3.3}$$

where the (pseudo)metrics[3] are defined as $\|h - g\|_{L_2(\mathbb{P}_{\mathcal{S}})} = \Big(\sum_{x \in \mathcal{S}} \frac{1}{n}\big(h(\boldsymbol{x}) - g(\boldsymbol{x})\big)^2\Big)^{\frac{1}{2}}$ and $\|h - g\|_{L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})} = \Big(\sum_{\boldsymbol{x} \in \widehat{\mathcal{S}}_{\mathcal{G},\epsilon}} \frac{p(\boldsymbol{x})}{n}\big(h(\boldsymbol{x}) - g(\boldsymbol{x})\big)^2\Big)^{\frac{1}{2}}, \forall h, g \in \mathcal{H}^{\mathsf{Lip}}$.

Remark 3.5. Theorem 3.4 shows that the previous covering number of $\mathcal{H}^{\mathsf{Lip}}$ evaluated at the sample $\mathcal{S}$ is upper-bounded by a new covering number of $\mathcal{H}^{\mathsf{Lip}}$ evaluated at any sample cover $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$, under a modified metric and at the cost of an additional error term that depends on ϵ. The equality trivially holds by taking $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon} = \mathcal{S}$, and by searching over all sample covers with different scale ϵ we may tighten the covering number for the invariant Lipschitz model class. Previous results on the covering number require $N\big(\tau, \mathcal{H}, L_2(\mathbb{P}_{\mathcal{C}})\big)/n \to 0$ as $n \to \infty$ so as to be non-vacuous (e.g., see [35]). Therefore, covering numbers depending on n is usually replaced by the one depending on the data dimension, which may lead to loose result when the data dimension is high (e.g., images). Now with the refined result in Theorem 3.4, covering numbers depending on n can potentially be replaced by the one depending on sample covering numbers $N\big(\tau, \mathcal{H}, L_2(\mathbb{P}_{\mathcal{C}})\big)$, thus steering away from the dependence on data dimension.

[3] The term $p(\boldsymbol{x})$ in defining the function norm $L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})$ can be viewed as the probability mass at $\boldsymbol{x}$ which equals to the number of examples that $\boldsymbol{x}$ covers. We refer the precise definition of $p(\boldsymbol{x})$ to the Appendix A

Theorem 3.4 directly leads to a refined empirical Rademacher complexity bound (and a refined generalization bound, see Appendix A). The following proposition is a weaker but more convenient version. We can set $\epsilon \to 0$ as $n \to \infty$ to further suppress the additional error term on large samples.

Proposition 3.6 (Refined Rademacher complexity of $\mathcal{G}$-invariant $\mathcal{H}^{\mathsf{Lip}}$)**.** Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathcal{H}^{\mathsf{Lip}}$ be a hypothesis class such that every $h \in \mathcal{H}^{\mathsf{Lip}}$ is κ-Lipschitz with respect to $\|\cdot\|$ and invariant to $\mathcal{G}$. Given any $\epsilon > 0$, $\alpha > 0$, let $\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}$ be an ϵ-cover of $\mathcal{S}$. Then

$$\mathfrak{R}_S(\mathcal{H}) \leq 4\kappa\epsilon\sqrt{1 - \frac{|\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}|}{n}} + 4\alpha + 12\int_{\alpha}^{\infty}\sqrt{\frac{\log N\big(\tau, \mathcal{H}, L_2(\mathbb{P}_{\widehat{\mathcal{S}}_{\mathcal{G},\epsilon}})\big)}{n}}d\tau. \tag{3.4}$$

3.2 Special Case: Linear Model Class

This subsection shows a more interpretable generalization benefit of model invariance by considering linear model class and linear data transformations (e.g., rotation). The following theorem provides a refined model complexity result for the invariant Linear model class.

Theorem 3.7 (Refined Rademacher complexity of $\mathbf{A}$-invariant $\mathcal{H}^{\mathsf{Linear}}$)**.** Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. Let $\mathbf{A}$ be the matrix representation of any linear data transformation. Consider the L_p-norm-bounded linear model class $\mathcal{H} = \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x}\rangle : \boldsymbol{w} \in \mathbb{R}^d, \|\boldsymbol{w}\|_p \leq W\}$ for some $p \geq 1$ and constant $W > 0$. Let $\mathcal{H}^{\mathsf{Linear}} = \{h \in \mathcal{H} : h(\boldsymbol{x}) = h(\mathbf{A}\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^d\}$ be the subset of $\mathcal{H}$ that is invariant under transformation $\mathbf{A}$. Then

$$\mathfrak{R}_S(\mathcal{H}^{\mathsf{Linear}}) = \frac{W}{n}E_\sigma\left[\inf_{\boldsymbol{\eta}\in\mathbb{R}^d}\|\boldsymbol{u}_\sigma + (\mathbf{A}-\mathbf{I})\boldsymbol{\eta}\|_q\right], \tag{3.5}$$

where $\boldsymbol{u}_\sigma = \sum_{i=1}^n \sigma_i \boldsymbol{x}_i$ and $\{\sigma_1, \ldots, \sigma_n\}$ are i.i.d. Rademacher random variables.

As a comparison, the Rademacher complexity of the general linear model class $\mathcal{H}$ is $\mathfrak{R}_S(\mathcal{H}) = \frac{W}{n}E_\sigma\left[\|\boldsymbol{u}_\sigma\|_q\right]$. Note that we always have the model complexity gap $\mathfrak{R}_S(\mathcal{H}) - \mathfrak{R}_S(\mathcal{H}^{\mathsf{Linear}}) \geq 0$ in Theorem 3.7 (as one can check by taking $\boldsymbol{\eta} = \mathbf{0}$ in (3.5)) and the gap can also be made strict in many cases. The following proposition gives a more interpretable result by further considering the L_2-norm-bounded linear model class.

Proposition 3.8 (Refined Rademacher complexity of L_2-norm-bounded $\mathbf{A}$-invariant $\mathcal{H}^{\mathsf{Linear}}$)**.** Let $\mathcal{H}^{\mathsf{Linear}}$ be the L_2-norm-bounded linear model class that is invariant under transformation $\mathbf{A}$ for some constant $W > 0$ (i.e., $p = 2$ in Theorem 3.7). Then

$$\mathfrak{R}_S(\mathcal{H}^{\mathsf{Linear}}) = \frac{W}{n}E_\sigma\left[\|\mathbf{P}\boldsymbol{u}_\sigma\|_2\right], \tag{3.6}$$

where $\mathbf{P} = \mathbf{I} - (\mathbf{A}-\mathbf{I})(\mathbf{A}-\mathbf{I})^\dagger$ and $(\mathbf{A}-\mathbf{I})^\dagger$ is the Moore–Penrose inverse of $\mathbf{A}-\mathbf{I}$.

Remark 3.9. Proposition 3.8 shows that the improvement in model complexity (and thus the generalization bound) for linear invariant models depends both on the sample and on data transformations. The matrix $\mathbf{P}$ in (3.6) is essentially the orthogonal projection matrix that projects the weighted sum of data $\boldsymbol{u}_\sigma$ onto the null space of $(\mathbf{A}-\mathbf{I})^T$. Intuitively, the linear data transformation $\mathbf{A}$ separates each input $\boldsymbol{x}$ into two orthogonal components: $\mathbf{P}\boldsymbol{x}$ that is $\mathbf{A}$-invariant, and $\boldsymbol{x} - \mathbf{P}\boldsymbol{x}$ that is $\mathbf{A}$-variant. Linear models that are invariant to $\mathbf{A}$ will ignore the $\mathbf{A}$-variant component and only capture the $\mathbf{A}$-invariant component (otherwise they will not be $\mathbf{A}$-invariant). Suppose that the data distribution has zero mean bounded variance, then the Rademacher complexity of $\mathcal{H}^{\mathsf{Linear}}$ is upper-bounded by the variance of the $\mathbf{A}$-invariant component in $\boldsymbol{x}$. Therefore, if the data transformation captures most of the data variance, the corresponding invariant models will have much smaller model complexity and thus better generalization performance. We give some examples in Example 3.10.

Example 3.10. Suppose the data have Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$. Let $\mathcal{H}$ be the L_2-norm-bounded linear model class. Then we have the following Rademacher complexity [22] bounds: (a) $\mathfrak{R}_n(\mathcal{H}) \leq \frac{W\sigma}{\sqrt{n}}$ for the general $\mathcal{H}$; (b) $\mathfrak{R}_n(\mathcal{H}') \leq \frac{W\sigma}{2\sqrt{n}}$ for the flipping-invariant $\mathcal{H}' \in \mathcal{H}$; (c) $\mathfrak{R}_n(\mathcal{H}'') \leq \frac{W\sigma}{n}$ for the circular-translation-invariant $\mathcal{H}'' \in \mathcal{H}$. Note that the convergence rate of $O(\frac{1}{n})$ in the Rademacher complexity of circular-translation-invariant $\mathcal{H}''$ is typically viewed as a fast speed that guarantees small generalization gap.

4 Data-driven Selection of Data Transformations

In this section, we discuss the data transformation selection problem to guide the development of model invariance. Given a task, we propose to evaluate the generalization benefit of a set of data transformations using its induced sample covering number. This model-agnostic evaluation is generally applicable and does not require constructing or training an invariant model beforehand. Then, we propose an algorithm to estimate the sample covering number.

4.1 A Suitability Measurement of Data Transformations

The pool of candidate data transformations for a given task may be infinitely large. Due to the lack of guidance, we usually make selections based on heuristics. In Section 3, we show that an invariant model is likely to generalize well if the corresponding data transformations induce a small sample covering number. Therefore, we propose to use sample covering number as an empirical suitability measurement of a set of data transformations.

Suitability Measurement. For a set of data transformations $\mathcal{G}$ and a dataset $\mathcal{S}$, we use the sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by $\mathcal{G}$ as the suitability measurement of $\mathcal{G}$.

Note that we use generalization benefit to indicate a small generalization gap. Since we also need a small training error (empirical risk) to guarantee a small generalization error, the proposed suitability measurement does not always indicate small generalization error. Nonetheless, we can choose data transformations by prior knowledge in practice to ensure that they preserve the label distribution and do not negatively impact the training error. This suitability measurement of data transformation does not need a model to be trained beforehand and is generally applicable to all models since it excludes the effect of model's implicit bias. Another advantage is that we can empirically evaluate and compare the potential generalization benefits of being invariant to different data transformations before model training which provide guidance on transformation selection.

4.2 Empirical Estimation of Sample Covering Numbers

Exactly computing the sample covering number is intractable in general since computing the transformation induced metric can be difficult and finding the smallest sample cover is NP-hard. To address this, we propose an algorithm, Algorithm 1, to estimate the sample covering number. For detailed steps (including Algorithm 2 which uses a clustering method to approximate the sample covering number) and computational complexity, please refer to Appendix C.

Algorithm 1 Empirical Estimation of Sample Covering Number

Input: Sample $\mathcal{S}$, a set of transformations $\mathcal{G}$, scale ϵ

Output: $\widehat{N}(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$, an empirical estimation of sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$

Algorithm:

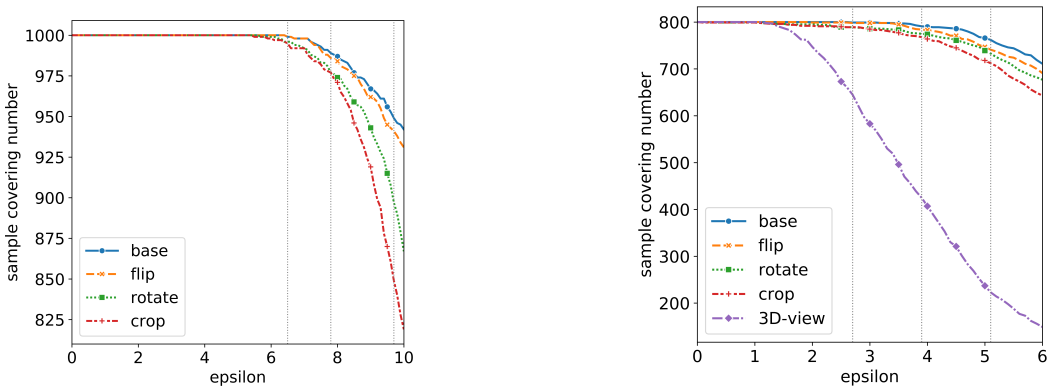
Step 1: Compute (or approximate) the direct orbit distance between any two examples $\boldsymbol{x}_s, \boldsymbol{x}_t \in \mathcal{S}$.

$$d_{\mathcal{O}}(\boldsymbol{x}_s, \boldsymbol{x}_t) = \|\mathcal{O}_{\mathcal{G}}(\boldsymbol{x}_s) - \mathcal{O}_{\mathcal{G}}(\boldsymbol{x}_t)\| = \min_{g_1, g_2 \in \mathcal{G}} \|g_1(\boldsymbol{x}_s) - g_2(\boldsymbol{x}_t)\|.$$

Step2: Compute the $\rho_{\mathcal{G}}$ distance between any two examples $\boldsymbol{x}_s, \boldsymbol{x}_t \in \mathcal{S}$ by solving a shortest path problem:

$$\rho_{\mathcal{G}}(\boldsymbol{x}_s, \boldsymbol{x}_t) = \min_{path(s,t)} \sum_{(i,j) \in path(s,t)} d_{\mathcal{O}}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Step3: Construct the pairwise distance matrix $\mathbf{D} \leftarrow [\rho_{\mathcal{G}}(\boldsymbol{x}_s, \boldsymbol{x}_t)]_{s,t}$ and approximate the sample covering number $\widehat{N}(\epsilon, \mathcal{S}, \rho_{\mathcal{G}}) \leftarrow$ `Distance2SampleCoveringNum`($\mathbf{D}$,ϵ) using Algorithm 2.



(a) CIFAR-10 (b) ShapeNet

Figure 1: Estimated sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by different data transformations at different scales ϵ. "base" indicates no transformation. Note that as ϵ increases, it starts to exceed the L_2 distance between some images and thus some images get covered by others without doing any transformation. Three vertical dashed lines indicate the maximum scale ϵ at which the "base" yields a certain sample covering number, and from left to right they are $100\%n$, $99\%n$, $95\%n$, respectively.

5 Experiments

In this section, we empirically verify whether our proposed sample covering number induced by data transformations correlates the actual generalization benefit for invariant models. Specifically, we first measure the sample covering number induced by different types of data transformations using the estimation Algorithm 1. Then, we investigate the actual generalization benefit for models invariant to those transformations in classification tasks and analyze the correlation.

Datasets. We conduct experiments on two public datasets, CIFAR-10 [20] and ShapeNet [8]. ShapeNet is a large-scale 3D data repository which enables us to do more complex data transformations (e.g., change of 3D-view) beyond commonly used 2D geometric transformations. In our experiment, we use a subset of ShapeNet which contains 10 categories. [10] provide, for each 3D object, a set of 24 multi-view pre-rendered images, and we use them to approximate the random perturbation of 3D-view. Appendix D provides more details about datasets and experimental settings.

Data Transformations. We choose some commonly used geometric transformations with typical settings (which are assumed to be label-preserving). On CIFAR-10, we choose *flipping*, *cropping*, and *rotation*. On ShapeNet, we choose *flipping*, *cropping*, *rotation*, and additionally the *3D-view* change. We choose the same type of data transformations with the same setting during estimating the sample covering number and evaluating the generalization benefit.

5.1 Estimation of Sample Covering Numbers

We apply Algorithm 1 to estimate the sample covering number induced by different data transformations. For efficiency, we randomly choose 1000 training images on CIFAR-10 and randomly choose 800 training images on ShapeNet. We choose L_2 norm as the input space norm used in Definition 3.2 which measures the distance between different images. For continuous data transformations, we do random sampling to approximate the orbit of an data point. Appendix D provides more estimation results with different sample sizes.

Sample covering number. Figure 1 illustrates the estimated sample covering numbers induced by different transformations at different scales ϵ. As the scale ϵ increases, the sample covering number $N(\epsilon, \mathcal{S}, \rho_{\mathcal{G}})$ induced by any data transformation starts to decrease, indicating a smaller-sized sample cover needed to cover the entire dataset. Meanwhile, different transformations behave differently. On CIFAR-10, cropping induces the smallest sample covering number. On ShapeNet, 3D-view transformation induces the smallest sample covering number and the gap is significant. Notably,

Model	$n = 100$		$n = 1000$		$n = all$	
	acc (%)	gap	acc (%)	gap	acc (%)	gap
Base	41.05 ± 0.52	58.95 ± 0.52	68.62 ± 0.90	31.38 ± 0.90	85.43 ± 0.35	14.57 ± 0.35
Flip	44.19 ± 0.74	55.81 ± 0.74	75.12 ± 0.20	24.88 ± 0.20	89.67 ± 0.24	10.33 ± 0.24
Rotate	47.02 ± 0.46	52.93 ± 0.51	76.07 ± 0.28	23.92 ± 0.27	89.91 ± 0.13	10.05 ± 0.16
Crop	$\mathbf{50.47 \pm 0.48}$	$\mathbf{49.53 \pm 0.48}$	$\mathbf{81.84 \pm 0.12}$	$\mathbf{18.15 \pm 0.11}$	$\mathbf{92.52 \pm 0.08}$	$\mathbf{7.48 \pm 0.08}$

Table 1: Classification accuracy and generalization gap (the difference between training and test accuracy) for ResNet18 on CIFAR10. n denotes the sample size per class.

Model	$n = 100$		$n = 1000$		$n = all$	
	acc (%)	gap	acc (%)	gap	acc (%)	gap
Base	67.75 ± 2.02	32.25 ± 2.02	83.33 ± 0.38	16.67 ± 0.38	91.81 ± 0.22	8.18 ± 0.22
Flip	69.75 ± 1.55	30.25 ± 1.55	84.24 ± 0.30	15.76 ± 0.30	92.07 ± 0.20	7.92 ± 0.20
Rotate	70.25 ± 1.19	29.50 ± 1.15	83.93 ± 0.38	15.94 ± 0.35	91.85 ± 0.20	8.03 ± 0.26
Crop	74.88 ± 1.03	23.53 ± 1.30	86.13 ± 0.39	13.75 ± 0.32	92.64 ± 0.12	7.17 ± 0.19
3D-View	$\mathbf{78.13 \pm 1.31}$	$\mathbf{14.94 \pm 1.76}$	$\mathbf{88.79 \pm 0.34}$	$\mathbf{8.38 \pm 0.79}$	$\mathbf{94.38 \pm 0.08}$	$\mathbf{3.09 \pm 0.10}$

Table 2: Classification accuracy and generalization gap (the difference between training and test accuracy) for ResNet18 on ShapeNet. n denotes the sample size per class.

3D-view transformation can cover the dataset ($n = 1000$) with 80% of the data while keeping ϵ below the minimum distance between any two images. According to our propositions, data transformations that are able to induce a smaller sample covering number would benefit more for the corresponding invariant models. Therefore, Figure 1 indicates that models should generalize well if it is invariant to 3D-view transformation on ShapeNet or cropping on CIFAR-10.

5.2 Evaluation of Generalization Benefit

To see how data transformations benefit generalization and to verify whether sample covering number is a good indicator of the generalization benefit, we learn corresponding invariant models for classification tasks and analyze the actual generalization benefit. We use ResNet18 [18] on both datasets and discuss the influence of model class's implicic bias in Appendix D. A simple method to learn invariant models is to do data augmentation. The augmented loss function is $\mathcal{L}_{aug}(\boldsymbol{x}) = \mathcal{L}(f(g(\boldsymbol{x})))$, where $f(\cdot)$ denotes the model and $g(\boldsymbol{x})$ denotes a randomly sampled example in $\boldsymbol{x}$'s orbit induced by transformation $\mathcal{G}$. We use this method on CIFAR-10 and ShapeNet and show results in table 1 and 2.

Sample covering number correlates well with generalization benefit. We use the generalization gap (the gap between training accuracy and test accuracy) to measure actual generalization benefit. Compared with the baseline, invariant models show an improved reduced generalization gap and also improved test accuracy. On CIFAR-10, cropping-invariant model shows the smallest generalization gap and the highest accuracy. On ShapeNet, the model that is invariant to 3D-view changes shows the smallest generalization gap and the highest accuracy, especially when the training data size is small. By comparing results in figure 1 and table 1-2, we observe a clear correlation between the smaller sample covering number and better generalization benefit. This verifies our proposition — invariance to more suitable data transformations gives the model more generalization benefit.

Model invariance indeed improves after learning. We conduct another experiment on ShapeNet with 3D-view transformation to verify that the improved generalization is indeed brought by model invariance. Instead of using data augmentation, we train models with an invariance regularization loss similar to [34]: $\mathcal{L}_{reg} = \mathcal{L}_{cls} + \mathcal{L}_{inv} = \mathcal{L}(f(\boldsymbol{x})) + \lambda \mathrm{KL}(f(\boldsymbol{x}), f(g(\boldsymbol{x})))$. Specifically, in addition to minimize the classification loss on the original image, we penalize the model by minimizing the KL divergence between model outputs of the original image and the transformed one. By adjusting λ, we can control the invariance extent. At test time, we use an empirical approximation of $\mathcal{L}_{inv}(\boldsymbol{x}) = \mathbb{E}_{g_1, g_2 \in \mathcal{G}}[\mathrm{KL}(f(g_1(\boldsymbol{x})), f(g_2(\boldsymbol{x})))]$ to evaluate model invariance under transformation

λ	train acc (%)	test acc (%)	gap	$\mathcal{L}_{inv}$	$\mathcal{A}_{inv}$
0	99.99 ± 0.01	91.81 ± 0.22	8.19 ± 0.22	0.0548 ± 0.0028	0.62 ± 0.00
0.01	99.98 ± 0.00	92.77 ± 0.16	7.21 ± 0.16	0.0290 ± 0.0029	0.74 ± 0.01
0.1	99.99 ± 0.00	93.87 ± 0.19	6.11 ± 0.19	0.0152 ± 0.0003	0.83 ± 0.01
0.3	99.98 ± 0.00	94.23 ± 0.11	5.76 ± 0.11	0.0117 ± 0.0004	0.85 ± 0.00
1	99.58 ± 0.04	94.68 ± 0.09	4.90 ± 0.09	0.0095 ± 0.0001	0.87 ± 0.00
3	97.74 ± 0.19	94.48 ± 0.19	3.26 ± 0.09	0.0060 ± 0.0003	0.88 ± 0.00
10	95.67 ± 0.26	93.56 ± 0.29	2.11 ± 0.04	0.0037 ± 0.0002	0.89 ± 0.00
100	72.04 ± 8.35	72.23 ± 8.10	0.19 ± 0.37	0.0004 ± 0.0001	0.82 ± 0.00

Table 3: Evaluation of ResNet18 on ShapeNet under 3D-view transformations. $\mathcal{L}_{inv}$ denotes the test invariance loss. $\mathcal{A}_{inv}$ denotes the test consistency accuracy (indicating whether model's prediction is unchanged after data transformation) under worst-case transformations.

$\mathcal{G}$. In Table 3, as we increase the invariance penalty by increasing λ, invariant models show smaller generalization gap. Moreover, the decreased invariance loss and increased consistency accuracy verify that the model invariance indeed improves after learning, and thus suggests that the generalization benefit is brought by model invariance.

6 Related Work

Input space perspective. The most related line of work is the perspective that characterizes the input space. [3, 4] show that the invariant representations can equivalently reduce the input dimension for downstream tasks and thus significantly reduce the model complexity (which is exponentially dependent on input dimensions) of downstream linear models. [29] factorizes the input space into the product of a base space and a finite set of data transformations. Since the covering number needed to cover the base space is smaller, they show that it leads to a reduced generalization bound for large margin classifiers. [27] also factorizes the input space using finite data transformations and show the reduced covering number results in improved generalization bound for invariant models. Compared with these works, our work tries to cover the sample instead of the input space which circumvents the exponential dependence on input dimensions and enables empirical evaluation.

Function space perspective. Another line of work directly characterizes the function space of invariant models. [21] uses PAC-Bayes to show the reduction of generalization upper bound when the model class is symmetrized to be invariant. [15] analyzes the function space under feature averaging operator and shows the first strict generalization gap (instead of upper bound) via a linear model. So far this line of work only considers the model invariance obtained exclusively via feature averaging.

Others. We mention some work that addresses the model invariance but does not focus on understanding the benefit. [1] mentions that the VC-dimension of an invariant model cannot be larger than its counterpart. [7] characterizes the general functional representations of invariant probability distributions and neural network structures that implement them. [9] uses group theory to characterize the benefit of learning with data-augmented loss.

7 Conclusion

This paper understands the generalization benefit of model invariance from a data perspective. By introducing sample cover induced by data transformations, we characterize the properties of data transformations and relates it to the improved generalization bound for invariant models. Our theoretical and empirical results motivate the exploration of more suitable data transformations on a given task and develop the corresponding invariant models for better generalization performance. One limitation of our work is that our model-agnostic way does not take into account the implicit bias of certain model classes, and thus cannot exactly predict the generalization benefit an invariant model would have. Besides, we do not have an extensive discussion on balancing the trade-off between improved generalization gap and reduced training accuracy. In future work, we would try to consider the implicit bias of certain model classes to better understand the benefit of model invariance.

References

[1] Yaser S. Abu-Mostafa. Hints and the VC Dimension. *Neural Comput.*, 5(2):278–288, March 1993. Place: Cambridge, MA, USA Publisher: MIT Press.

[2] Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations in hierarchical architectures. *arXiv:1311.4158 [cs]*, Mar 2014. arXiv: 1311.4158.

[3] Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised Learning of Invariant Representations in Hierarchical Architectures. *arXiv:1311.4158 [cs]*, March 2014. arXiv: 1311.4158.

[4] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On Invariance and Selectivity in Representation Learning. *arXiv:1503.05938 [cs]*, March 2015. arXiv: 1503.05938.

[5] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial Learning Guarantees for Linear Hypotheses and Neural Networks. *arXiv:2004.13617 [cs, stat]*, April 2020. arXiv: 2004.13617.

[6] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.

[7] Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *arXiv:1901.06082 [cs, stat]*, September 2020. arXiv: 1901.06082.

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[9] Shuxiao Chen, Edgar Dobriban, and Jane Lee. A Group-Theoretic Framework for Data Augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21321–21333. Curran Associates, Inc., 2020.

[10] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 628–644. Springer, 2016.

[11] Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, June 2016. ISSN: 1938-7228.

[12] Taco S. Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, page 9142–9153, 2019.

[13] R. M. Dudley. A course on empirical processes. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XII - 1982*, page 1–142. Springer Berlin Heidelberg, 1984.

[14] Shimon Edelman. Class similarity and viewpoint invariance in the recognition of 3d objects. *Biological Cybernetics*, 72(3):207–220, Feb 1995.

[15] Bryn Elesedy and Sheheryar Zaidi. Provably Strict Generalisation Benefit for Equivariant Models. *arXiv:2102.10333 [cs, stat]*, February 2021. arXiv: 2102.10333.

[16] Robert Gens and Pedro M. Domingos. Deep symmetry networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *NIPS*, page 2537–2545, 2014.

[17] Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific Reports*, 10(1):1411, Jan 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[19] Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 2752–2765. Association for Computational Linguistics, Jul 2020.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the Benefits of Invariance in Neural Networks. *arXiv:2005.00178 [cs, stat]*, April 2020. arXiv: 2005.00178.

[22] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[23] Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. Learning with group invariant features: A kernel perspective. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[24] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[25] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL (1)*, page 5582–5591. Association for Computational Linguistics, 2019.

[26] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[27] Akiyoshi Sannai and Masaaki Imaizumi. Improved Generalization Bound of Group Invariant / Equivariant Deep Networks via Quotient Feature Space. *arXiv:1910.06552 [cs, stat]*, March 2020. arXiv: 1910.06552.

[28] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially Robust Generalization Requires More Data. In *Advances in Neural Information Processing Systems 31*, pages 5014–5026. Curran Associates, Inc., 2018.

[29] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Generalization Error of Invariant Classifiers. *arXiv:1610.04574 [cs, stat]*, July 2017. arXiv: 1610.04574.

[30] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[31] Marysia Winkels and Taco S. Cohen. 3d g-cnns for pulmonary nodule detection. *CoRR*, abs/1804.04656, 2018.

[32] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher Complexity for Adversarially Robust Generalization. volume 97 of *Proceedings of Machine Learning Research*, pages 7085–7094, Long Beach, California, USA, June 2019. PMLR.

[33] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, page 3391–3401, 2017.

[34] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Learning Representations*, 2019.

[35] Tong Zhang. Covering number bounds of certain regularized linear function classes. *J. Mach. Learn. Res.*, 2:527–550, Mar 2002.

[36] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NIPS*, page 649–657, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]